+

Machine Translation and Its Role in Professional Translation

Špela Vintar, Dept. of Translation Studies
University of Ljubljana
Spela.vintar@ff.uni-lj.si

Outline

- Myths about MT
- How does MT work
  - Overview of approaches
  - Understanding data-driven MT
- MT in practice
- MT and professional translation
- Post-editing MT
- Further reading

Myths about MT

1. **MT does not work – machines will never be able to translate literary texts.**

   According to estimates, 90 % of translated texts involve documentation projects, technical specifications, user manuals & similar.

2. **MT is not for professional translators, only for amateurs**.

   Most CAT tools now integrate MT as a standard component when no fuzzy match was found.
   MT is an integral part of most localization projects.

+

Myths about MT

3. **MT is cheap, a translation is produced at the click of a button.**

   The development of rule-based systems is extremely costly, and the development of data-driven systems needs massive amounts of data.

4. **MT will never be available for small language pairs.**

   With growing availability of digital resources it is now possible to train SMT systems for many marginal language pairs.

Myths about MT

5. **Only an MT system can produce a translation as ridiculous as …**

   Think of bad human translations …

6. **In future there will be no need for human translators.**

   The translator's job description might change, but there will always be need for tech-savvy language professionals, and the global demand for language services is still growing.

Knowing your enemy...

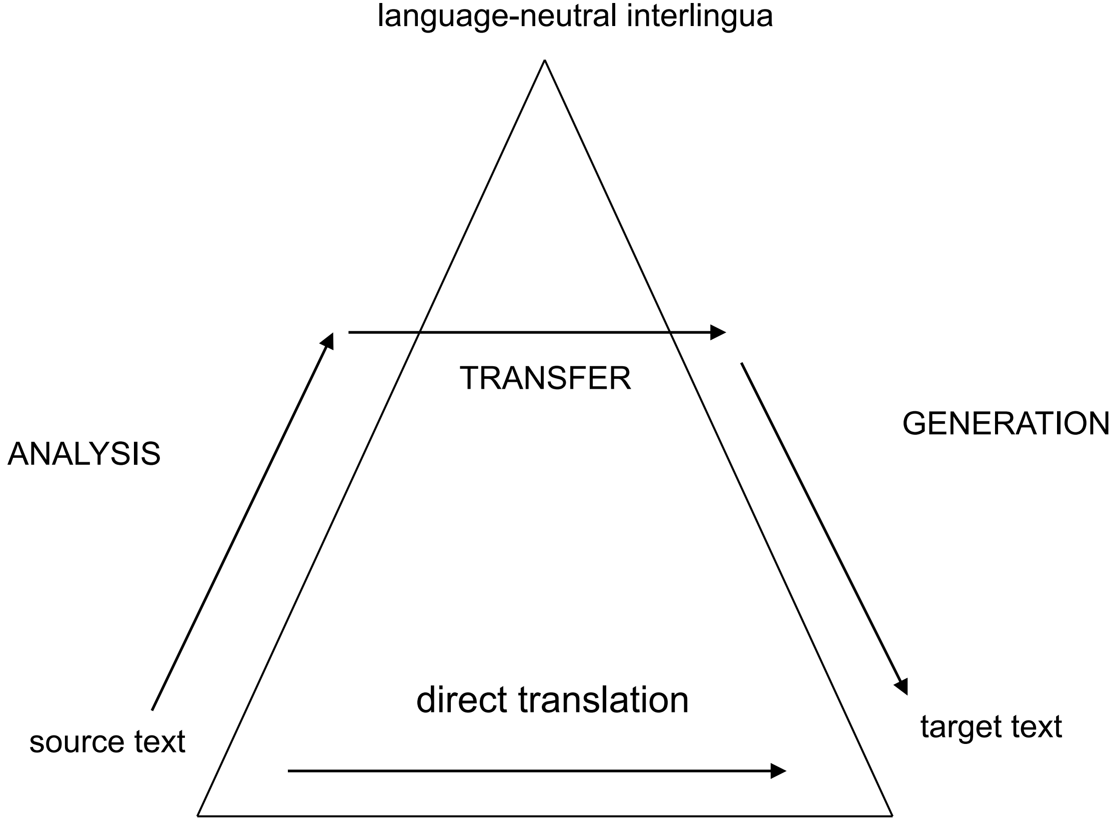
Historically, several approaches to Machine Translation have evolved:

- Direct
- Transfer
- *Interlingua*
- Statistical
- Example-based

Rule-based approaches

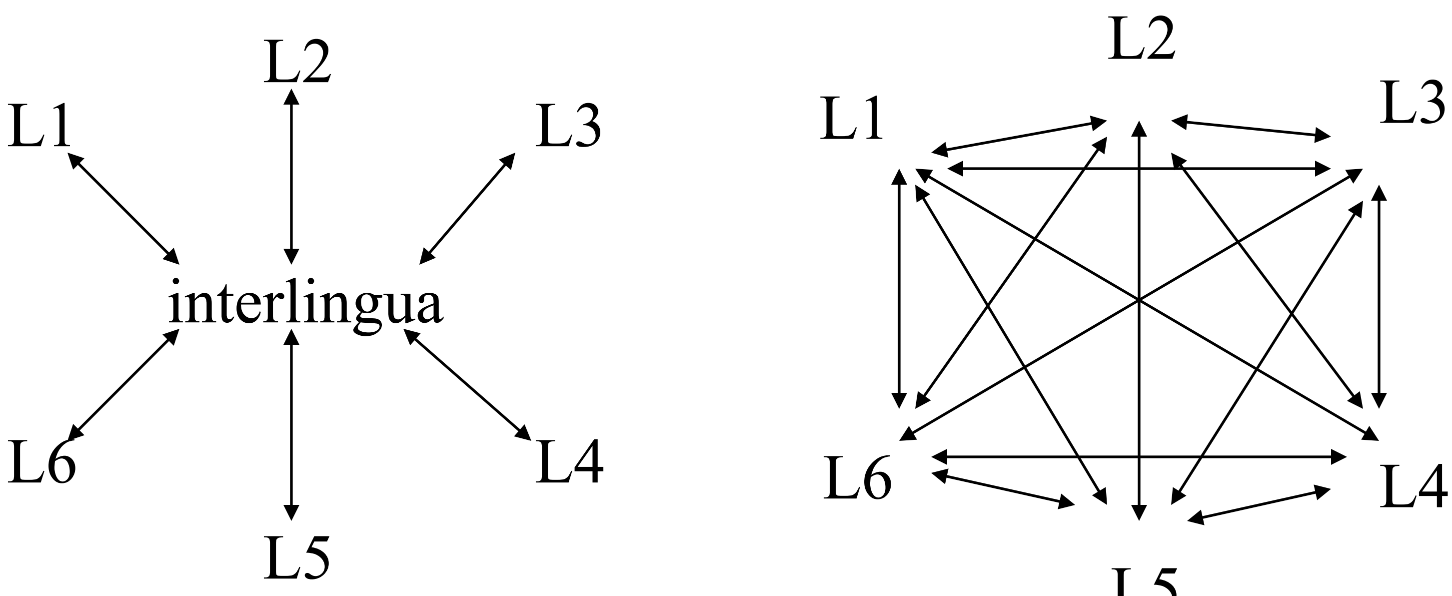
Data-driven approaches

Rule-based approaches & level of abstraction



The Vauquois Pyramid

Interlingua vs. transfer



Sources of translation errors

- Lexicon
  - Several possible equivalents for a source word, lexical gaps etc.
- Syntax
  - Large syntactic transformations and changes in word order
- Syntactic-semantic
  - No similar structure available in target language for the proposition to be translated
- Idiomaticity, Phraseology, Collocations

Lexicon

- Ambiguous source word with different equivalents for each sense
  - *financial* bank vs. *river* bank
  - In the *course* of things... (9 senses)
  - Time flies like an arrow.

- Lexical gaps: source word denotes a sense which is not lexicalized in the target language
  - eg. *cozy*

- Target language has a more fine-grained structure of senses than source language
  - eg.: EN *wall* -> DE *Wand* (internal), *Mauer* (external)

+ Syntactic-semantic transformations

- Change of structure:
  I like swimming
  "Ich schwimme gern"
  I swim gladly

- Verb complements:
  Jones likes the film.
  "Le film plait à Jones."

- Passive constructions:
  - eg. French-> English
    Ces livres se lisent facilement
    *"These books read themselves easily"
    These books are easily read

Idiomaticity

- Non-compositional meaning
- Examples:
  - George is a bull in a china shop.
  - He kicked the bucket.

SI: George je bik v trgovini s porcelanom. Je brcnil vedro.

HR: George je bik u staklarni. On je udario kantu.

FI: George on norsu posliinikaupassa. Hän potkaisi ämpäri.

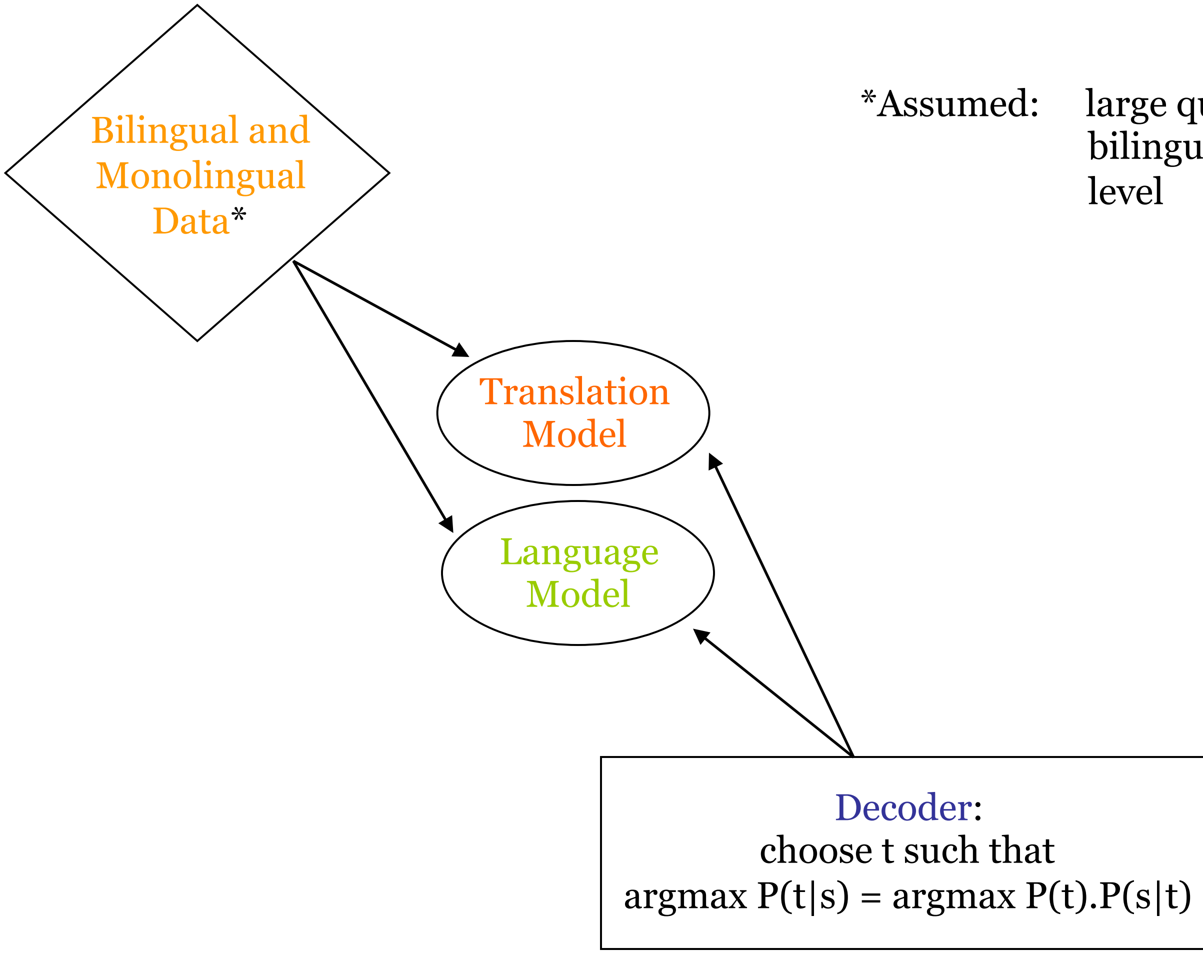
ES: George es un toro en una cacharrería. Dio una patada al cubo.

GR: Ο Γιώργος είναι ένας ταύρος σε υαλοπωλείο. Κλώτσησε τον κάδο.

Do we need rules?

- Children learn language without rules…
- …so why not machines?
- From 1990s, increasing amounts of digital texts available on the web, including parallel texts.
- Basic idea of SMT:
  - If we have a text and its translation, and we perform sentence-alignment:
    - For each source word in the source sentence the system can check whether there is a target word that continuously occurs in the translated sentence.

Statistical Machine Translation (SMT)



*Assumed: large quantities of high-quality bilingual data aligned at sentence level

Source: Andy Way. 2011, SMT Walkthrough

+

Language Modeling

- A language model assigns a probability to *every* string in that language.
- In practice, we gather a huge database of utterances and then calculate the relative frequencies of each.
- Problems?
  - many (nearly all) strings will receive *no* probability as we haven't seen them …
  - all unseen good and bad strings are deemed equally unlikely …
- Solution: How do we know if a new utterance is valid or not? By breaking it down into substrings -> **n-gram models**

The Translation Model

the language model

SMT:argmax $P(e|f) = \text{argmax } P(e) . P(f|e)$

the translation model

<u>Remember:</u>

If we carry out, for example, French-to-English translation, then we will have:

- an English Language Model, and
- an English-to-French Translation Model.

When we see a French string f, we want to reason backwards ...
What English string e is:

- likely to be uttered?
- likely to then translate to f?

We are looking for the English string e that maximises

$$P(e) * P(f|e).$$

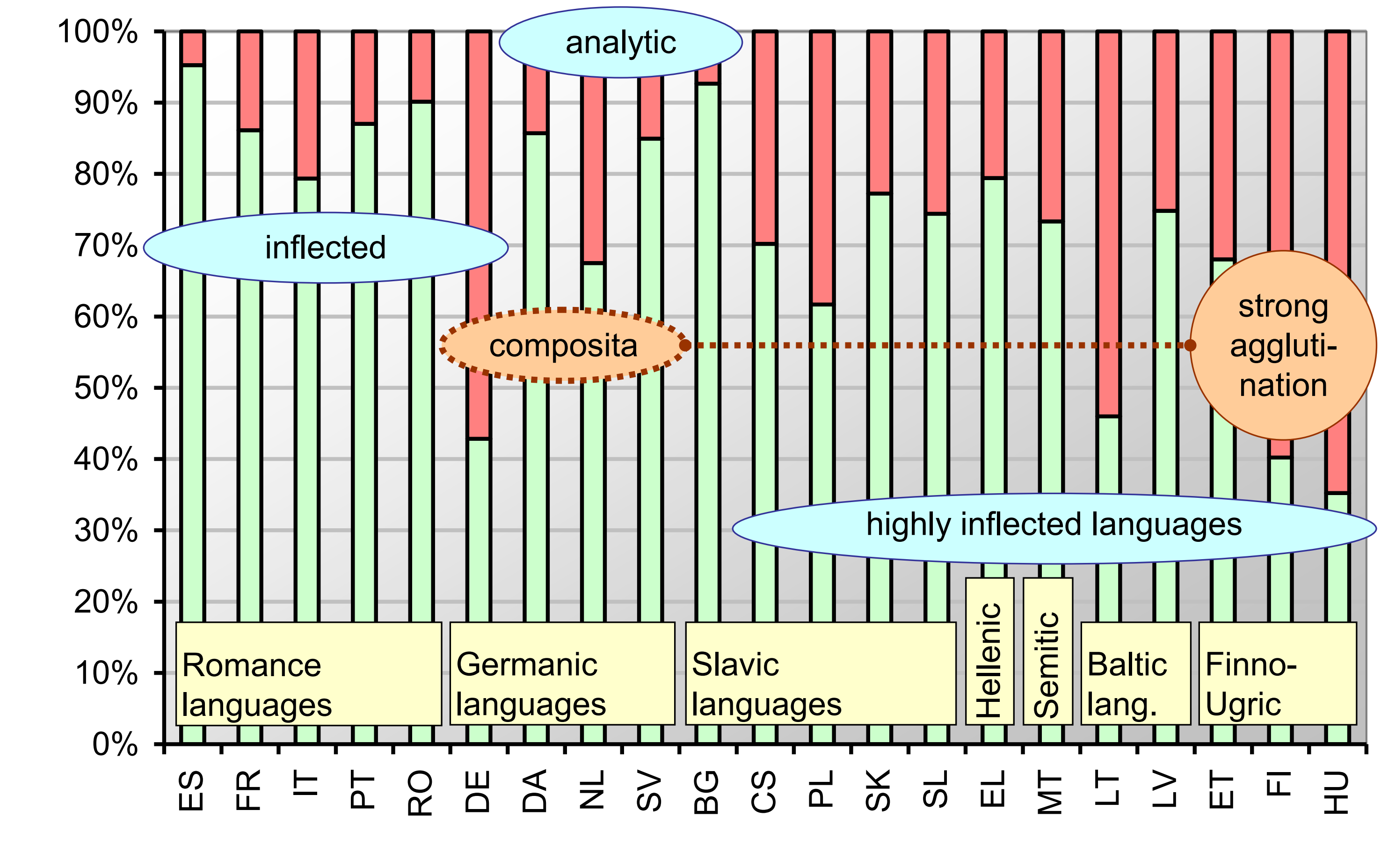
Statistical Machine Translation: Milestones

- In 2000, Egypt / Giza SMT toolkits freely available for research purposes
- Google's first MT attempts in 2002.
- In 2004, Franz Josef Och joins Google and soon starts heading their language research unit.
- Today, 57 supported languages.
- In 2009, Microsoft launches Bing, their own web-based free translation service.
- Moses: an EC-funded free SMT engine.

SMT and RBMT in practice

- Question is often not “HT or MT”, but “Either MT or No translation at all”.
- MT@EC: the EU is gradually switching to MT and retraining their staff to post-edit MT output.
  - EuroMatrix and EuroMatrix Plus projects have developed SMT for all 23 languages (using English as pivot).
  - In order to translate 6,8 mio pages -> 8500 full-time translators would be needed.
- Large corporations use MT (Adobe, Microsoft, Volkswagen, SAP …).
- Contrary to common belief, the contest between SMT and RBMT is not over yet…

+



DGT's SMT *maturity check* outcome as a (useful useless) sentences ratio + morphology

MT evaluation

- What is a good human translation?
  - Subjectivity
  - Purpose of the translation

- MT evaluation:
  - Back-translation
    HR: Pas je išao preko ceste.
    EN: The dog walked across the road.
    HR: Pas hodao preko ceste.
  - Human
  - Automatic

Human evaluation

- Human:
  - Various methods proposed….
  - Most widely accepted method by ARPA:
    - **Fluency**: rating how good the target language is.
    - **Adequacy**: rating how much information is transferred between the original and the translation.

Automatic evaluation

- A good metric will correlate closely to human judgement
- Metrics usually measure the similarity between MT output and a human (reference) translation
- Several metrics proposed:
  - BLEU
  - NIST
  - Meteor
  - WER
  - …

MT and professional translation

What translators say:

- Using MT is worse and slower than having to translate from scratch (D).
- Translating with the help of MT is of course faster (D).
- Post-editing MT output is more difficult and less predictable (D).
- Post-editing MT output is easier and more predictable (RU).
- Never had to do it and hope I never will (EN).
- It can be useful in many situations (NO).
- It hinders the natural flow of typing – more clicking, copy-pasting. It's physically tiring (D, F).
- It's the only way of matching clients' expectations regarding turnaround (F, RU).

(Source: Weiss, I. (2011) MT Oslo)

MT and professional translation

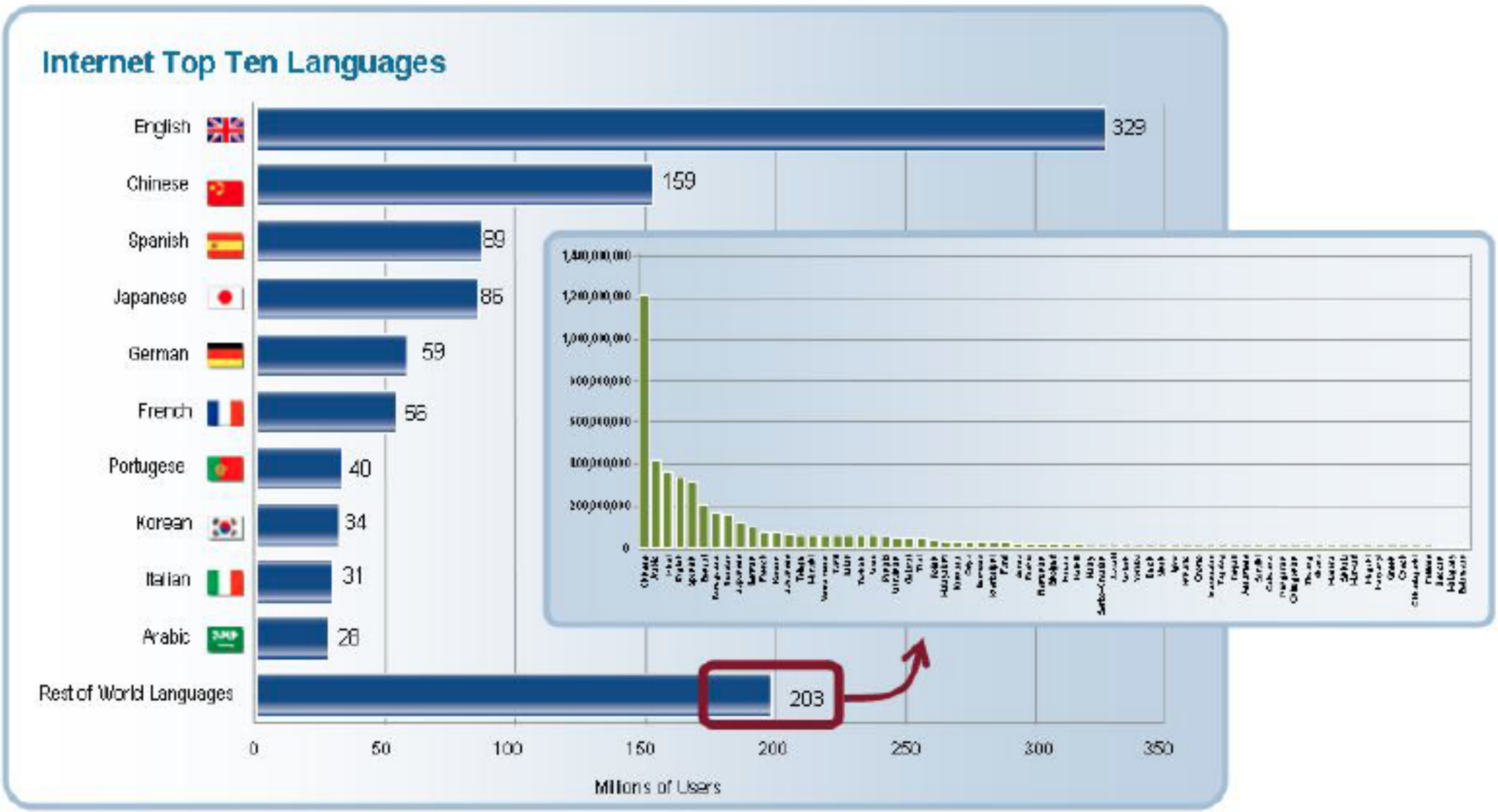
What translation vendors say:

- Deploy and customize MT to start translating more for less!
- Saving Time, Lowering Costs, Improving Service...
- In 2010, PEMT represented 200k out of 20M words (1%).
  In 2012, we expect a total volume of over 30M words, with at least 7M post-edited (23,3%).

What clients say:

- Does the translation industry have a "cost disease"? (No gains in productivity when considering translation proper)
- Quality is what each customer says it is.
- How much quality is enough?

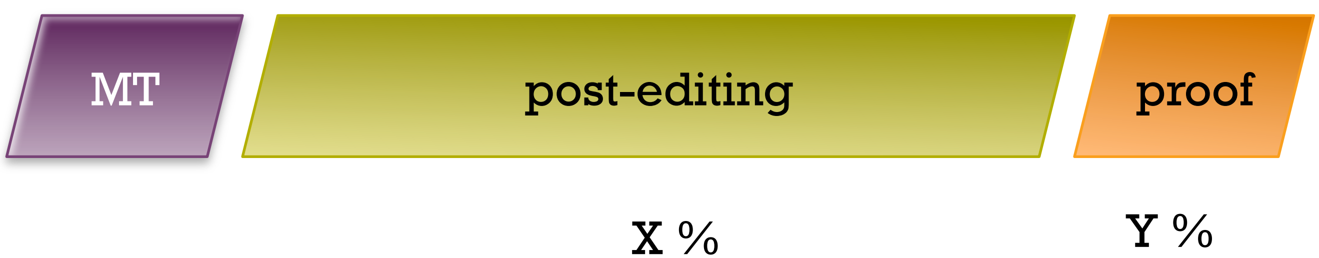
The long tail of languages on the web



New model of user-defined quality

Traditional:

Scalable:

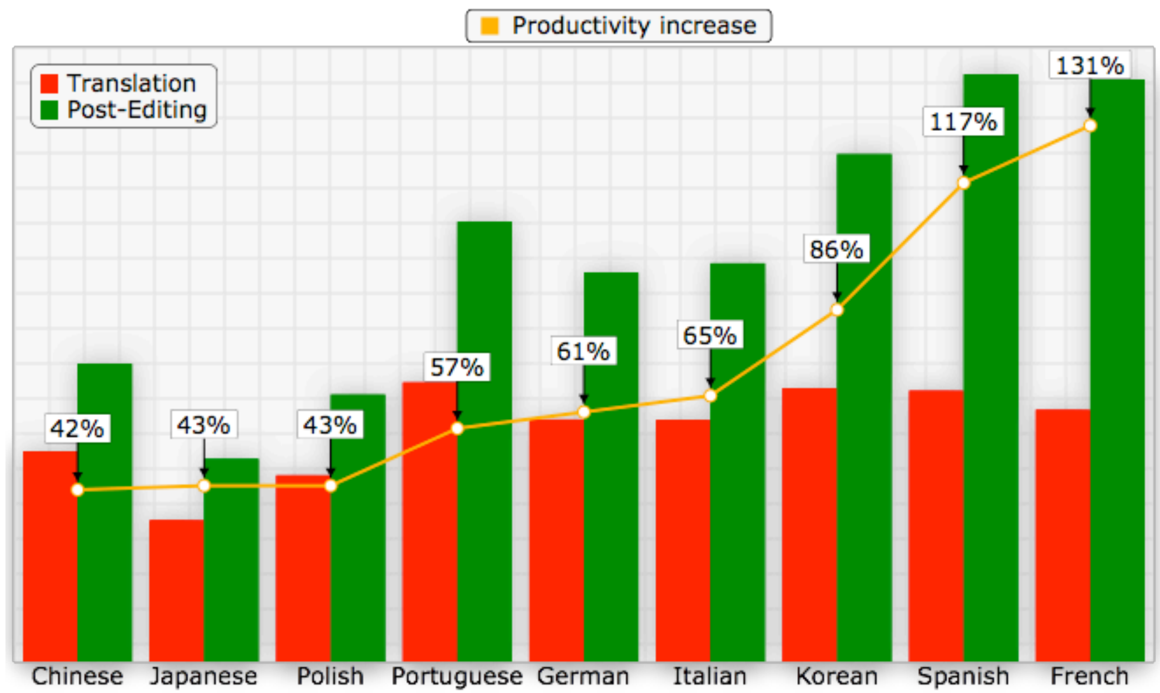


Post-editing MT

- …yes, there is an abbreviation: PEMT
- MT is regularly used in localization projects where no TM match is found.
- 2006: O'Brien establishes a correlation between MT and TM segments in the 80-90 fuzzy match range.
  - Dangerous! Many IPMs still use this rate to pay for PE.
- Translators generally resent doing it….
- …..but do it more and more often.

Post-editing MT: Productivity gain

- In most studies, PEMT outperforms translating from scratch (productivity gains from 42 to 131%).
- For some language-pairs, PEMT outperforms fuzzy matches.



Source: MT at Autodesk, http://translate.autodesk.com

Post-editing MT: TM vs MT

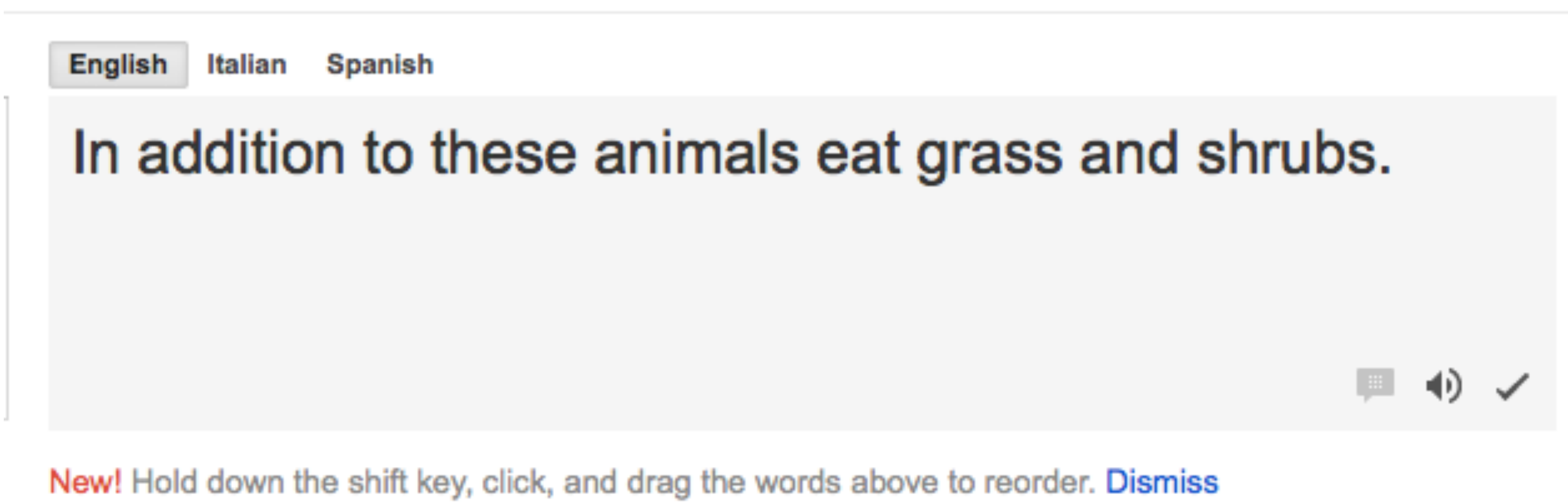
- Interesting study by Guerberof, A. (2009)...
- Comparing productivity and quality on a group of 8 professional translators, giving them New, TM and MT segments to translate
- Results:
  - Translators have higher productivity and quality when using machine-translated output than when processing fuzzy matches from TMs.
    - -> possible explanation: if the text is running smoothly, translators overlook terminology inconsistencies etc. and do not consult the original as often as they should.

MT and Professional Translation

- You are likely to be required to post-edit MT output and produce translations of various quality levels.
  - "gisting" vs. "raw translation" vs. "without grammatical errors" vs. "indistinguishable from HT"
  - MT output needs getting used to...
  - Post-editing MT is not necessarily done by translators.
  - Post-editing effort must be measured and evaluated for each individual project!
- You are likely to work with hybrid CAT/MT tools.
  - Editing a fuzzy match is not so much different from editing MT output. -> Typology of errors / corrections?
  - Domain-adapted MT systems perform much better than Google Translate.

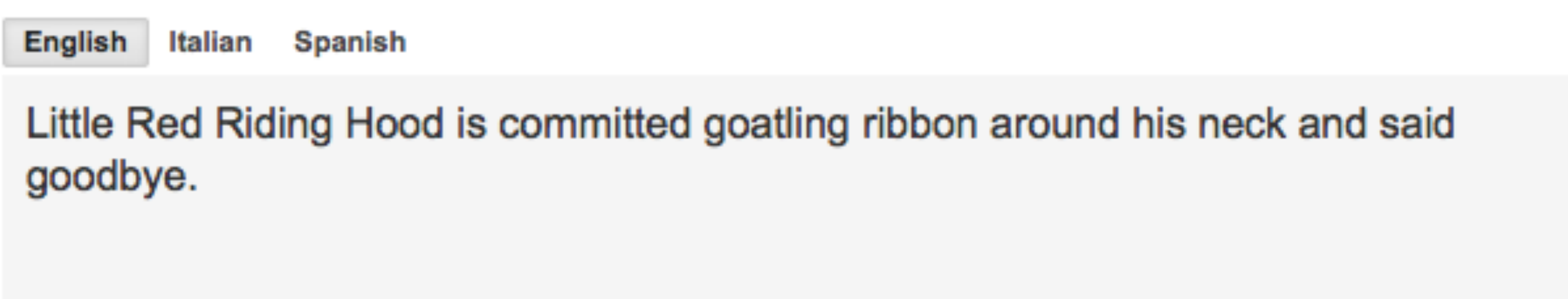
+ MT and Professional Translation

- Your students most likely already use Google Translate and/or other web services.
  - "If you can't beat it, eat it!" Practice using MT as an instant memory aid, discuss MT errors with students, play around with Google's alternative translations and "shift-and-drag" reordering.



MT and Professional Translation

- Your students most likely already use Google Translate and/or other web services.
  - "If you can't beat it, eat it!" Practice using MT as an instant memory aid, discuss MT errors with students, play around with Google's alternative translations and "shift-and-drag" reordering.
  - Discuss the lexical choices of MT systems and offer explanations for them.
    - Most MT systems are black box, but some insight into their strategies is useful.



MT and Professional Translation

- You most likely already use Google Translate and/or other web services.
  - "If you can't beat it, eat it!" Practice using MT as an instant memory aid, discuss MT errors with students, play around with Google's alternative translations and "shift-and-drag" reordering.
  - Discuss the lexical choices of MT systems and offer explanations for them.
    - Most MT systems are black box, but some insight into their strategies is useful.
  - Compare the outputs of different MT systems, for different text types.
    - If feasible, compare language pairs; translating the same content from different source languages.
  - Play around with back-translation.

MT and Professional Translation

- You are not likely to lose your job because of MT.
  - ...but knowing, understanding and participating in the development of MT systems is already a competitive advantage.

References

- Weiss, Isabelle, 2011. Machine Translation. Myths, Anecdotes and Realities. http://*www.**translator**portalen.com/.../TD-2011-**Machine_Translation**_1.ppt*
- Guerberof, Ana. Productivity and quality in MT post-editing. http://*www.mt-archive.info/MTS-2009-**Guerberof**.pdf*
- *Eisele, Andreas, 2010. Machine Translation at the European Commision. Presentation at the conference "10 Jahre Verbmobil".*
- The Future of Localization (2011). Lionbridge White Paper. http://en-us.lionbridge.com/translation-localization/knowledge-center/future-localization.htm
- MT Archive, a comprehensive bibliography of MT and related disciplines, http://www.mt-archive.info/
- Statistical Machine Traslation, http://www.statmt.org/
- EAMT, http://www.eamt.org/index.php
- Hutchins, J. W. and Somers, H. (1992) An introduction to machine translation. London: Academic Press, http://www.hutchinsweb.me.uk/IntroMT-TOC.htm